平台下粒子滤波结合改进 ABC 算法的 IoT 大数据特征选择方法 *

吴　颖 [1]，李晓玲 [1]，唐晶磊 [2]

(1. 中原工学院 信息商务学院，郑州 451191; 2. 西北农林科技大学 信息工程学院，陕西 杨凌 712100)

摘　要：针对现有物联网大数据特征选择算法计算效率低下、可扩展性不高的问题，提出一种基于改进人工蜂群(ABC)选择特征的系统架构，该架构包含四层体系，可以高效地聚合有效数据，剔除不需要的数据。整个系统是基于 Hadoop 平台、MapReduce 以及改进 ABC 算法的。改进 ABC 算法用于选择特征，而 MapReduce 则由并行算法支持，该算法可高效处理大量数据集。该系统使用 MapReduce 工具实现，并利用粒子滤波来消除噪声。将所提出的算法与同类方法进行比较，并通过使用十个不同的数据集对效率、准确性和吞吐量进行评估。结果表明，相比其他几种较新的算法，提出的算法在选择特征时更具可扩展性和高效性。

关键词：物联网；大数据；人工蜂群算法；特征选择；粒子滤波；小生境技术

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.04.0287

Internet of things big data feature selection method based on particle filter and improved ABC algorithm on Hadoop platform

Wu Ying[1], Li Xiaoling[1], Tang Jinglei[2]

(1. *College of Information & Bossiness, Zhongyuan Institute of Technology, Zhengzhou* 451191, *China*; 2. *School of Information Engineering, Northwest A&F University, Yangling Shaanxi* 712100, *China*)

Abstract: Aiming at the problem that the existing Internet of things big data feature selection algorithm has low computational efficiency and low scalability, this paper proposed a system architecture that selects features by using improved artificial bee colony. The architecture included a four-layer system and it could efficiently aggregate the effective data and eliminate unwanted data. The entire system was based on the Hadoop platform, MapReduce, and improved ABC algorithms. The method used improved ABC algorithm to select features and it also used a parallel algorithm to support MapReduce, which could efficiently process a huge volume of data sets. It used MapReduce tool to implement the system and it used particle filter for removal of noise. Compare the proposed algorithm with similar algorithms and evaluate the efficiency, accuracy and throughput by using ten different data sets. The results show that the proposed algorithm is more scalable and efficient in selecting features.

Key words: Internet of things; big data; artificial bee colony algorithm; feature selection; particle filter; niche technology

0　引言

物联网(IoT)是连接物理世界和网络世界的纽带。物联网技术的进步使得新型应用和服务配置需求高的物理世界的数字化，各种各样的东西在互联网的帮助下被分组在一起以共享信息。利用 IoT 可以感知物理环境，收集数据，传输或传播数据，处理适当应用程序的数据以及与其他事物进行通信，给人们的生活带来很大的方便。

但在实施方面，IoT 带来了非常大的挑战[1]。由于 IoT 是异构事物的混合体，与传统网络有很大不同，其可扩展性变得更加复杂[2]。此外，在 IoT 中相互通信的设备会消耗大量的内存和带宽。因此，物联网倾向于生成大量数据，称为大数据。为了应对这些限制，理想的解决方案就是绿色物联网。通过开展环境监测来减少排放和污染，以降低运营成本以及功耗[3]。在目前的大数据情况下，数据库供应商已经引入了各种标准和平台用于数据聚合以及数据分析。但这些平台通常功能单一，无法在 IoT 大数据中广泛使用。

基于以上分析，特征选择是处理大数据的核心方法之一。特征选择包括图像分类、聚类分析、数据挖掘、模式识别和图像检索等[4]。特征选择算法分为两大类：滤波方法(filter)和包装(wrapper)方法。在基于滤波的技术中，将为每个要素计算权重值，以便可以选择具有更好值的要素来表示原始大数据集。包

收稿日期：2018-04-20；**修回日期：**2018-05-28　**基金项目：**国家自然科学基金面上项目（61472314）

作者简介：吴颖（1987-），女，河南南阳人，讲师，硕士，主要研究方向为物联网、大数据等；李晓玲（1982-），女，河南博爱人，副教授，硕士，主要研究方向为物联网、计算机应用；唐晶磊（1974-），男，河北邢台人，副教授，博士，主要研究方向为大数据、智能算法等.

装技术则利用特征的子集来产生一组提名特征。之后，使用准确性来评估特征集的结果，它能取得比前者更好的结果。此外，蚁群优化[5]、粒子群优化算法[6]、蝙蝠算法[7]和人工蜂群(Artificial Bee Colony, ABC)[8,9]也被提出来以提高计算效率。现有特征选择算法存在很多缺点，比如实时连续数据难以提取特征，并且使用传统工具来处理大量数据时效率低下。

本文提出了一个系统架构，用于聚合大数据，利用改进ABC 算法选择特征，并将数据转发到 Hadoop 平台进行并行处理。

1 相关研究

特征选择是一个选择特征子集的过程，可以运用搜索技术遍历空间以实现特征选择，但这种方法对于识别大量特征似乎不切实际。于是，科研人员想到将群体智能技术和神经网络技术用于特征选择。同样的，Hadoop 分布式文件服务器也可用于特征选择，该服务器在计算机节点上具有多个本地磁盘，从而提供更好的数据局部性[10]。在具有高性能计算集群的系统中，计算机节点连接到一个名为 Lustre 的并行文件系统。Lustre 提供了一个高效且可扩展的数据存储设施。

Lustre 系统安装在使用 Lustre 作为本地存储的群集上。这些本地存储适用于传统的 MapReduce 功能，这些功能可以分两步完成，即读写操作。由于 Lustre 系统的读写吞吐量很高，这些操作提供了高速数据路径。Lustre 内部传输所需的时间取决于许多因素，如集群互联、数据加载等，这些因素在组合时会对传统的 MapReduce 功能产生影响。

MapReduce 编程可以生成大量的数据集，这些数据集对应现实世界任务的广泛多样性。MapReduce 将输入数据分成完全并行处理的小独立块。 MapReduce 体系结构将映射输出分类并发送到 reduce 作业，任务的输入和输出保存在文件系统中。Google 文件系统就是受到 MapReduce 模型的启发，它利用大量的计算机集群，通过交换机以太网进行联合。Google MapReduce 方案降低了广泛分布式应用程序的机群成本。MapReduce 方法使建立过程更简单、更容易。它基于实时执行，并没有定义节点的预先计划执行调度。MapReduce 范例可以在分布式节点上执行。MapReduce 模型可以获得较强的容错能力并且平衡每个集群的负载，使科研人员可以更轻松简单的执行操作。Google 的 MapReduce 结构最初是分布式文件系统，用于标识数据的位置和可访问性。

粒子群优化技术也可用于特征选择和处理大型数据集，使用该算法能够降低系统的复杂性，提高效率。从大型数据集中提取特征时会需要更多的时间，在这个过程中，使用不同的噪声数据过滤算法会对提取的特征子集产生重要影响。

基于上述卷积方法和传统 Hadoop 技术的方法需要系统从大量数据中选择最优的特征。为此，本文提出了一种基于Hadoop 平台和改进 ABC 算法的物联网大数据特征选择方法。改进 ABC 算法可以有效地选择最优功能，而 Hadoop 生态系统与改进 ABC 结合可以产生最佳结果。

2 提出的方法

2.1 四层分层结构

本文方法包括四层体系结构模型，每一层都有不同的功能支持，使读写操作能够高效运行，如图 1 所示。设计的模型可以帮助不同的对象使用共享的媒介进行交互。所提出的体系结构模型可以在应用程序生成各种差异数据。

第一层 通过各种对象生成、处理数据，然后收集和聚合数据。由于生成数据涉及不同数量的对象，因此，整个过程会以各种格式，不同的起源点为基础，周期性生成大量的异构数据。而且，各种数据都有安全性、隐私和质量的要求。此外，在传感器的数据中，元数据始终大于实际测量值。因此，在该层应用了早期注册和过滤技术，以过滤不必要的元数据以及冗余数据。

第二层 该层为各种设备提供端到端的连接。此外，在不同设备产生的数据汇集在这一层上，并以适当的形式进行排列。

第三层 特征提取和处理层是整个系统体系结构的主要层，它完成数据的特征提取和处理部分。由于本文需要实时数据流和离线数据分析，因此，需要一个第三方的实时工具与处理服务器相结合来提供实时的数据处理，可以使用 Strom、Spark、VoltDb 和 Hupa 作为该辅助工具。例如，可以使用MapReduce 来实现数据分析，使用改进 ABC 算法能够使本文方案更好地从大型数据集中获取特征。在这一层， MapReduce使用了和 HDFS 相同的结构。有了这个系统，本文也可以使用HIVE、HBASE 和 SQL 来管理数据库来存储历史信息。

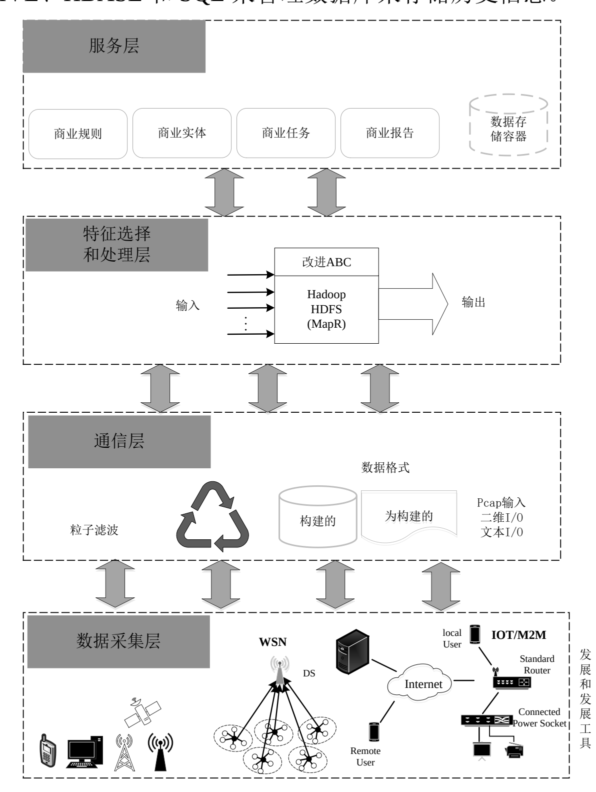


图 1 四层通信模型

第四层 服务层是负责将第三方接口合并到对象和人的最

底层。该层可以自主地用作单个站点，与其他位置合并或部署在云界面中。该层还能实现其他功能，例如，唯一的全局标识管理是在应用程序层中处理整个 Universe 中标识对象的关键元素。此外，本文提出的构建图层涉及到人类与各种智能对象的交互，因此，在应用层面需要一种智能算法，可以高效地与人进行交互。服务层的功能包括会话启动、设置通信规则、与异构对象交互以及终止会话等。

2.2　基于 Hadoop 和改进 ABC 的 HIABC 算法

为了详细阐述所提出的系统架构的体系结构，设置服务场景如图 2 所示，包括智能交通控制部门、智能天气预报部门以及智能医院和卫生部门。上述组件负责收集 IoT 网络中的异构数据，可以充当框架的底层。这些组件通过 GSM、Wi-Fi、3G 和 4G 等异构接入技术与智能决策和控制系统连接，智能决策系统位于智能城市框架的中间层次。

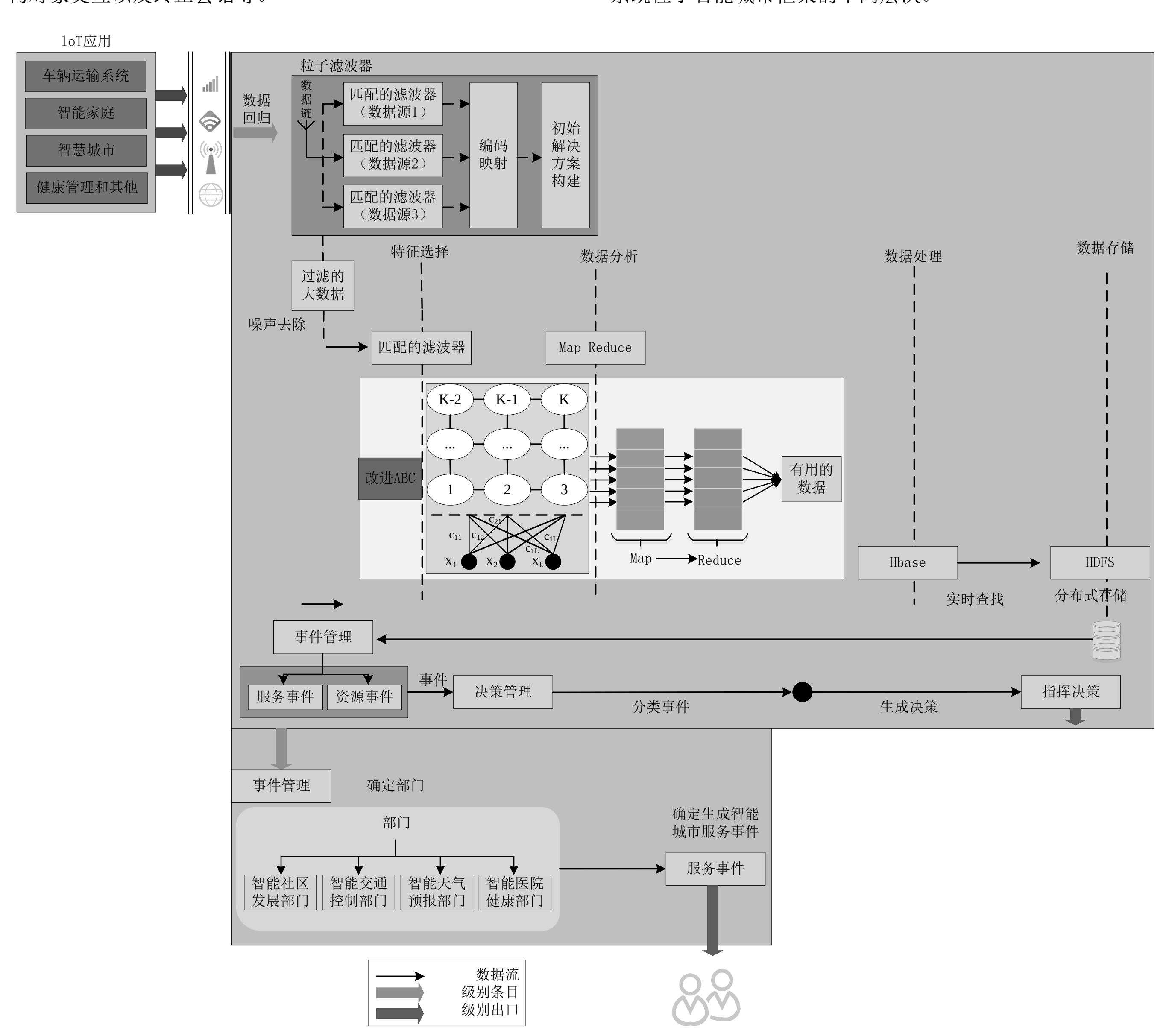


图 2　HIABC 系统体系结构

一个现实的 IoT 环境不仅包含大量的数据，还包含复杂的计算和多个应用程序[11]。IoT 系统的实现依赖于数据的获取和计算分析。智能环境理念旨在优化住宅资源、减少交通拥堵、提供有效的医疗服务。获取与日常运营活动相关的数据对于实现上述目标至关重要，但是，由于人员和其他连接设备产生大量数据，数据采集问题变得尤为艰难。因此，本文考虑将数据转换成数字数据。低成本和高能效的传感器已经成为从城市 IOT 获取异构数据的有效机制。随着连接设备数量的增加，城市变得更加智能[12]。因此，在城市郊区内部广泛部署异构传感器能够促进智能城市架构的形成。这些传感器负责收集来自邻近环境的不同类别的实时数据。

本文提出的方案的底层由多个组件组成。智能家居的关键是提高住宅建筑的能源利用率。家用电器配备了一个传感器，它决定了实时能源消耗，并将数据传送到中间层。数据处理层为特定家庭的能源消耗定义了一个阈值。数据过滤过程由数据聚合技术执行以确定超过阈值的值，从而进一步优化处理。车辆运输系统的主要目标是减少城市交通拥堵。数据处理级别定义了在两个规定点之间传输的平均时间。部署在路边的传感器收集车辆在两点之间的出入信息。嵌入式聚合技术通过分析规定位置的当前行程时间来确定拥堵道路。气象部门的作用是确

定天气条件和其他环境参数。例如，部署在某些位置的传感器可以监测城市的一氧化碳浓度。这些传感器将采集到的数据传送到中间层进行相应地过滤和处理，以便于决策和事件生成。

所提出的架构使用多种通信技术，包括 ZigBee、蓝牙、Wi-Fi 以及数据和蜂窝网络等把感测数据传输到数据处理层，以进行数据滤波、分析、处理、存储、决策等。所以，这一层被视为框架的大脑。为了执行上述任务，将多个模式嵌入到此层中。最初，大量的感测数据通过聚合机制进行过滤，以获得有价值的实时和离线数据。MapReduce 范例用于数据分析，而操作和存储由 Hadoop 分布式文件系统(HDFS、HBASE 和 HIVE 执行。

2.2.1 基于粒子滤波的数据过滤

聚合技术通过应用数据过滤来提高数据处理效率，在提出的框架中采用粒子滤波（particle filter，PF）[13]执行数据过滤。PF 是一个最优估计器，它可以从感测数据中去除噪声。PF 的主要思想是将后验概率密度函数用一组特殊的随机样本表示，以估计出不同状态的最小方差。

假设在 PF 中，用 $p(x_0)=p(x_0|z_0)$ 表示状态的初始概率密度函数，则根据贝叶斯理论可知，状态预测方程为

$$p(x_k|z_{1:k-1})=\int p(x_k|x_{k-1})p(x_{k-1}|z_{1:k-1})dx_{k-1} \tag{1}$$

状态更新方程为

$$p(x_k|z_{1:k})=\frac{p(z_k|x_k)p(x_k|z_{1:k-1})}{p(z_k|z_{1:k-1})} \tag{2}$$

式（2）中：

$$p(z_k|z_{1:k-1})=\int p(z_k|x_k)p(x_k|z_{1:k-1})dx_k \tag{3}$$

利用蒙特卡洛算法将整个计算过程简单化，即将上式中的积分计算过程离散化，将其变为对一组带有权值的样本求和。

令 $\pi(x_{0:k}|z_{0:k})$ 为重要性函数，它是由概率密度函数 $p(x_{0:k}|z_{0:k})$ 得到的，从 $\pi(x_{0:k}|z_{0:k})$ 中取出 N 个独立的样本：$\{x_{0:k}^i, i=1,2,\ldots,N\}$

可以得到状态的后验期望：

$$\left.\begin{aligned}\hat{E}[f_k(x_{0:k})]&=\sum_{i=1}^{N}f_k(x_{0:k}^{(i)})\tilde{w}_k^{(i)}\\ \tilde{w}_k^{(i)}&=\frac{w_k^{*(i)}}{\sum_{j=1}^{N}w_k^{*(i)}}\end{aligned}\right\} \tag{4}$$

其中：$w_k^{*(i)}=\dfrac{p(z_{0:k}|x_{0:k})p(x_{0:k})}{\pi(x_{0:k}|z_{0:k})}$ 表示没有归一化的重要性权值。

利用重采样方法解决计算过程中的粒子退化问题，即引入有效粒子个数 N_{eff}：

$$N_{eff}=\frac{1}{\sum_{i=1}^{N}(\tilde{w}_t^{(i)})^2} \tag{5}$$

当 $N_{eff}<N_{th}$ 时，系统进行重采样，否则的话，进行下一步计算。N_{th} 表示阈值，由实际情况而定。

2.2.2 数据的存储和处理

本文所提出的方案在 Hadoop 框架中存储和处理数据。利用 MapReduce 分析过滤数据[14]。MapReduce 分两步工作。首先是将过滤数据集转换为另一组数据的映射，然后将映射过程中创建的数据组合在一起，并生成一组数量减少的值。数据存储和处理在实现智慧城市中发挥着重要作用。如图 2 所示，所提出的框架利用多种技术，如 HDFS、HBase、HIVE 等来满足上述要求。HDFS 是 Hadoop 的主要存储空间，它的存储是分布式的，能够满足大数据处理的可扩展性需求。为了支持自主决策，整个集群上的实时读/写功能至关重要，因此，HBase 用于提高 Hadoop 的处理速度，因为它提供了实时查找内存中缓存的功能，此外，它还增强了系统的可用性和容错性。HIVE 通过驻留在 Hadoop 集群上的大量数据提供查询和管理功能。由于 SQL 不能用于查询 HIVE，本文使用 HiveQL 来查询 Hadoop 集群上的数据

2.2.3 HIABC 算法

本文提出了 HIABC 算法用于大数据集中的特征选择。人工蜂群算法（ABC）是一种随机搜索的元启发式全局优化算法，由三个部分组成：食物来源、雇佣的蜜蜂以及未雇佣的蜜蜂[15]。具体解释如下：

a）食物来源代表了给定问题的解决方案。

b）雇佣蜜蜂被用来找出不同的食物来源。此外，它们还用于存储信息，并与蜂窝中的其他蜜蜂共享此信息。

c）未雇佣的蜜蜂分为两类，即旁观者蜜蜂和侦察蜜蜂。旁观者蜜蜂收到来自被雇用蜜蜂的共享信息，这些信息用于寻找更好质量的食物来源；当被雇用的蜜蜂在寻找食物来源精疲力竭时，它们就变成了侦察蜜蜂，并试图寻找新的食物来源。

ABC 算法的主要过程如下所示：

1　初始化状态
2　执行
　　雇佣的蜜蜂

$$a_{ij}=a_j^{min}+rand(a_j^{\max}-a_j^{min}) \tag{6}$$

$i=1,2,3,\ldots,N$， $j=1,2,3,\ldots,K$，其中 N 和 K 是实物来源和优化参数。

未雇佣的蜜蜂

$$v_{ij}=x_{ij}\Phi(x_{ij}-x_{kj}) \tag{7}$$

v_i 为实物源， j 和 k 是随机变量。

$$fitness_i=\begin{bmatrix}\dfrac{1}{1+f_i} if\ f_i\ge 0\\ 1+abs(f_i) if\ f_i<0\end{bmatrix} \tag{8}$$

$$p_i=\frac{fitness_i}{\sum_{n=1}^{F} fitness_i} \tag{9}$$

3 记录最可能的结果

4 当周期达到最大时，结束。

在 ABC 算法的搜索过程中有可能会出现停滞现象，因此本文引入改进后的小生境技术[16]减少停滞现象的发生。如果两个侦察个体间的距离小于设定的阈值 L，则对它们之中适应度值小的个体进行惩罚，以增加它在后面的进化过程中被淘汰的概率。经过这一步骤后，优良个体就会分散在约束空间中，可以更好的保持种群多样性。即在算法中，当侦察蜂在子群内搜索时，如果在固定的进化代数内，适应度最高的两代个体满足：

$$\begin{aligned}&\left\|x^i-x^j\right\|\le L\\&\left|F(x^i)-F(x^j)\right|\le\theta\end{aligned} \tag{10}$$

就表明这个子群出现了停滞现象，因此需要把它淘汰，重新进行初始化。式（10）中， $2\le i\le l, 2\le j\le l$ ，且 $i\ne j$， $\left\|x^i-x^j\right\|$ 代表距离，L 代表设定的阈值，θ 代表子群个体适应度值的标准差，l 代表子群内的最大进化代数， $F(x^i),F(x^j)$ 表示子群第 i、j 代中的最优个体的适应度值。

在 HIABC 的一个特定场景中，每个食物来源与一个位矢量(大小为 N，其中 N 是特征的总数)相关联。矢量中的位置与需要评估的特征总数一致。在这种情况下，如果约定特征的值等于 1，则表明该特征是评估子集的一部分。如果特征值等于 0，则表明特征不是评估子集的一部分。 此外，食物来源储存其质量信息，可由分类器指定的特征子集的精度给出。

HIABC 算法用于特征选择的步骤如下：

a）在 Hadoop 处理系统中，当使用粒子滤波从数据集中去除噪声时，系统利用前向搜索策略[17]找到最佳和最低数量的特征。在前向搜索策略中， N 个食物来源含有 N 个特征。

b）每个食物来源的特征子集被分配给分类器，其中它使用准确度作为适度值(准确度被存储在食物源的适合度中)。

c）利用修正率参数(MR)来确定所选食物来源的邻居，并雇佣蜜蜂访问每个食物来源并探索邻居。为了提取特征，从最初食物来源的位向量创建邻居。在位矢量的每个位置生成一个随机的数字 R_i ($0<R_i<1$)。如果该值小于扰动参数 MR，则该特征被注入到子集中

d）如果新发现的食物来源的质量优于探索食物来源的质量，则认为邻近的食物来源是最新的食物来源。这些信息与其他蜜蜂共享。在 HIABC 中，数据的大小呈指数增长，所以这个过程一直持续到 Hadoop 中选择最佳参数为止。

e）执行基于小生境的种群淘汰机制，选择最佳个体。

f） N 个特征被随机创建并提交给分类器，新发现的来源被分配给侦察蜜蜂，并且雇佣蜜蜂再次执行它们的任务。

3 实验分析

对本文提出的体系结构进行测试，并与随机森林算法、文献[18，19]算法进行对比。对每种方法在相同的数据集上进行测试，总共进行十次重复实验，以实验结果的平均值作为最终结果。所有的实验都是在安装在 Ubuntu 14.04 LTS 中多集群的 Hadoop 上进行的。所提出的特征选择算法以 Java 编程语言实现。

3.1 数据集

本文提出的优化和特征选择方法在 UCI 机器学习库[20]的 10 个数据集上进行测试，使用多群集 Hadoop 系统在不同学习算法上测试每个数据集。每个数据集的描述如表 1 所示。每个数据集上主要是根据特征的数量对算法进行分析。

表 1 用于测试分析的数据集

数据集	特征数	示例数
室内活动监测传感器	11	928 438
空气质量	15	9251
GPS 轨迹	15	163
来自 RSS 的室内用户移动预测	4	13 397
3D 路况网络	4	454 865
污水处理厂	38	567
肝炎	19	155
住房	14	512
云	10	1 028
用于情绪分析的 Twitter 数据集	2	2 073

3.2 结果讨论

基于 IABC 的特征选择的性能评估和准确度是通过使用具有 K 个不同分区的 10 倍交叉验证来获得的。在这个过程中，其中一个分区用作主分区，其余 $K-1$ 个分区用作训练集，该过程重复十次，最终结果为所有十个分区的平均值。此外，在测试中建立的特征使用 Z 分机制进行了归一化处理，该机制将每个特征集的平均值相减，并将其除以该集的标准差。

UCI 数据集的不同特征影响了特征选择算法的性能和准确性。表 2 中给出了选定数量的特征和全部特征的准确性情况比较。表 3 是在相同条件下本文算法与随机森林算法、文献[18]算法以及文献[19]算法的分析比较结果，所选特征比其他原始特征列表小得多。与其他方法相比，在大多数的数据集中，本文提出的特征选择算法在准确性方面表现更好。在诸如空气质

量，3D 路况网络和云等数据集中，本文算法的准确性较差，而在某些情况下，如 GPS 轨迹和用于情绪分析的 Twitter 数据集中，本文算法的准确度几乎与其他方法相同，其余情况下，本文算法的准确度要优于另外三种算法。总体而言，本文提出的算法在准确性方面能够表现出较好的结果。

表 2　提出的系统在 UCI 数据集中的准确性

数据集	特征总数	选择的特征数	精度(%)	平均精度 (%)
室内活动监测传感器	12	5	73.35	94.36
空气质量	15	8	72.88	83.91
GPS 轨迹	15	5	64.64	77.02
来自 RSS 的室内用户移动预测	5	3	73.96	84.32
3D 路况网络	4	1	74.51	82.77
污水处理厂	39	15	80.73	95.67
肝炎	17	10	64.39	92.09
住房	18	6	56.65	81.87
云	10	6	74.73	78.25
用于情绪分析的 Twitter 数据集	2	1	83.56	91.53

表 3　本文方案与其他方法的准确性比较

数据集	随机森林算法 (%)	文献[18]算法 (%)	文献[19]算法 (%)	本文算法 (%)
室内活动监测传感器	84.83	79.26	77.95	94.36
空气质量	92.65	92.99	90.25	83.91
GPS 轨迹	70.88	68.72	66.72	77.02
RSS 用户移动预测	77.02	77.58	74.60	84.32
3D 路况网络	88.85	90.74	87.52	82.77
污水处理厂	89.27	87.66	82.63	95.67
肝炎	87.25	88.29	82.46	92.09
住房	66.89	74.68	76.11	81.87
云	83.32	85.31	72.37	78.25
用于情绪分析的 Twitter 数据集	90.16	88.79	89.16	91.53

图 3 为在 GPS 轨迹数据集下，选择不同的特征个数时，四种算法的分类准确性结果比较图。从图中可以看出，本文提出的方案在 GPS 轨迹数据集上的分类性能优于其余三种特征选择算法，并且随着特征个数的增加，分类准确性也在不断提高。当特征选择个数为 12 时，本文算法有最好的准确率 85.62%。

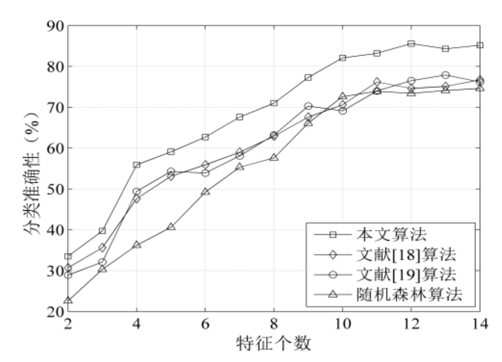


图 3 GPS 轨迹数据集分类准确性

接下来将基于 IABC 算法的系统与单个节点 Hadoop、基于 Java 的查询系统进行比较，结果如图 4 所示。在单个节点 Hadoop 中，每次处理数据时都没有任何系统优化。另一方面，基于 Java 查询的系统通过所有其他群优化方法进行测试。过滤系统用于在将数据传递到 Hadoop 生态系统之前从数据中去除噪音。随着本文数据量的逐渐增加，三种系统均能够实时处理大量数据，但相同数据量的情况下，本文系统的所需的处理时间明显比其余两种系统的处理时间少，本文系统能够实时有效地处理数据并生成帮助对象做出决策的结果。例如，实时处理环境数据有助于对象避免去那些污染严重的地方。

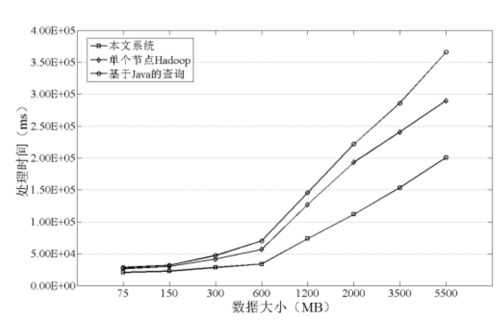


图 4 不同系统的处理时间分析比较

本文还通过增加数据的大小来测试提出系统的吞吐量。如图 5 所示，吞吐量与数据集的大小成正比。当数据集的大小增加时，吞吐量也会增加。最初，三种系统的数据处理速度相差并不大，但是，随着数据集大小的增加，单个节点 Hadoop 和基于 Java 的查询系统处理速度大大降低。相比之下，本文方案依旧能够保持较高的效率。

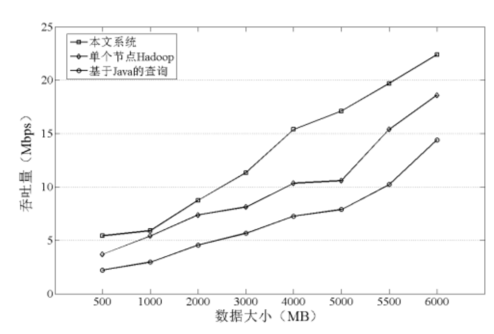


图 5 不同系统的吞吐量比较

为了用其他数据集进行测试和验证，本文测量了在不同医疗数据集上的处理时间，如图 6 和 7 所示。由图可知，本文所提出的方案需要几秒钟来处理以 GB 为单位的数据。此外，如果增加数据集的大小，吞吐量也会被最大化，因此，从这些结果可以得出结论，本文所提出的具有并行处理的系统具有更好的数据处理效果。

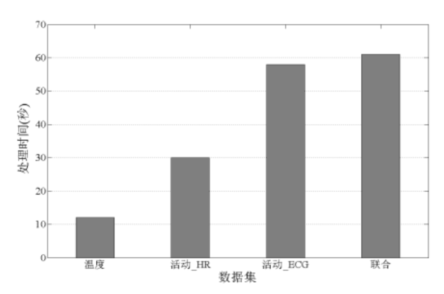


图 6 医疗数据集的处理时间

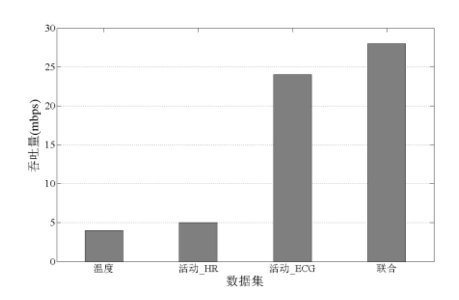


图 7 本文提出系统在不同数据集的吞吐量

4　结束语

本文提出了大数据物联网中特征选择的系统架构。所提出的方案基于四层体系结构模型，可以高效地聚合有效数据，剔除不需要的数据。整个系统利用改进蜂群算法选择特征，通过 Hadoop 生态系统处理数据，系统基于 MapReduce 工具实现，并利用粒子滤波来消除噪声。结果证明，在 Hadoop 生态系统中使用 IABC 可显着提高系统特征选择的效率。

参考文献：

[1] 赵会群，李会峰，刘金銮. RFID 物联网复杂事件模式聚类算法研究 [J]. 计算机应用研究，2018，35 (2): 339-341. (Zhao Huiqun, Li Huifeng, Liu Jinluan. Study on RFID complex event pattern clustering algorithm of Internet of things [J]. Application Research of Computers, 2018, 35 (2): 339-341.)

[2] 钱志鸿，王义君. 物联网技术与应用研究 [J]. 电子学报，2012，40 (5): 1023-1029. (Qian Zhihong, Wang Yijun. IoT Technology and Application [J]. Acta Electronica Sinica, 2012, 40 (5): 1023-1029.)

[3] Ahmad A, Paul A, Rathore M, *et al.* An efficient multidimensional big data fusion approach in machine-to-machine communication [J]. ACM Trans on Embedded Computing Systems, 2016, 15 (2): 32-39.

[4] 魏葆雅，林梦雷，郑艺峰. 基于标记重要性的多标记特征选择算法 [J]. 湘潭大学自科学报，2017，39 (4): 1-5. (Wei Baoya, Lin Menglei, Zheng Yifeng. Multi-label feature selection algorithm based on labeling-importance [J]. Natural Science Journal of Xiangtan University, 2017, 39 (4): 1-5.)

[5] 吕琳，尉永清，任敏，等. 基于蚁群优化算法的凝聚型层次聚类 [J]. 计算机应用研究，2017，34 (1): 114-117. (Lyu Lin, Wei Yongqing, Ren Min, *et al.* Agglomerative hierarchical clustering based on ant colony optimization algorithm [J]. Application Research of Computers, 2017, 34 (1): 114-117.)

[6] Tang Y, Guan X. Parameter estimation for time-delay chaotic system by particle swarm optimization [J]. Chaos Solitons & Fractals, 2017, 40 (3): 1391-1398.

[7] 尚俊娜，刘春菊，岳克强，等. 多智能体蝙蝠算法在无线传感器中的应用 [J]. 传感技术学报，2015，28 (9): 1418-1424. (Shang Junna, Liu Chunju, Yue Keqiang, *et al.* The multi-agent bat algorithm applied to wireless sensor networks [J]. Chinese Journal of Sensors and Actuators, 2015, 28 (9): 1418-1424.)

[8] Akay B, Karaboga D. A modified Artificial bee colony algorithm for real-parameter optimization [J]. Information Sciences, 2012, 192 (1): 120-142.

[9] Civicioglu P, Besdok E. A conceptual comparison of the Cuckoo-search, particle swarm optimization, differential evolution and artificial bee colony algorithms [J]. Artificial Intelligence Review, 2013, 39 (4): 315-346.

[10] Shvachko K, Kuang H, Radia S, *et al.* The Hadoop distributed file system [C]// Proc of IEEE Symposium on MASS Storage Systems and Technologies. Washington DC: IEEE Computer Society, 2010: 1-10.

[11] Rong W, Zhang X, Dave C, *et al.* Smart city architecture: a technology guide for implementation and design challenges [J]. China Commun, 2014, 11 (3): 56-69.

[12] Sanchez L, Muñoz L, Galache J A, *et al.* SmartSantander: IoT experimentation over a smart city testbed [J]. Computer Networks, 2014, 61 (6): 217-238.

[13] 崔丽珍，吴迪，赫佳星，等. 基于改进粒子滤波的井下跟踪算法研究与实现 [J]. 计算机应用研究，2017，34 (5): 1476-1479. (Cui Lizhen, Wu Di, He Jiaxing, *et al.* Research and implementation on underground tracking algorithm based on improved particle filter [J]. Application Research of Computers, 2017, 34 (5): 1476-1479.)

[14] Dean J, Ghemawat S. MapReduce: a flexible data processing tool [J]. Communications of the ACM, 2010, 53 (1): 72-77.

[15] Bao L, Zeng J C. Comparison and analysis of the selection mechanism in the artificial bee colony algorithm [C]// Proc of International Conference

on Hybrid Intelligent Systems. 2009: 411-416.

[16] Quabab B Y. Niching particle swarm optimization with local search for multi-modal optimization [J]. Information Sciences, 2012, 197 (197): 131-143.

[17] 王晓梅，林晓惠，黄鑫. 基于特征有效范围的前向特征选择及融合分类算法 [J]. 小型微型计算机系统，2016, 37 (6): 1159-1163. (Wang Xiaomei, Lin Xiaohui, Huang Xin. Algorithm of forward feature selection and aggregation of classifiers based on feature effective range [J]. Journal of Chinese Computer Systems, 2016, 37 (6): 1159-1163.)

[18] 姚登举，杨静，詹晓娟. 基于随机森林的特征选择算法 [J]. 吉林大学学报：工学版，2014, 44 (1): 137-141. (Yao Dengju, Yang Jing, Zhan Xiaojuan. Feature selection algorithm based on random forest [J]. Journal of Jilin University: Engineering and Technology Edition, 2014, 44 (1): 137-141.)

[19] 张进，丁胜，李波. 改进的基于粒子群优化的支持向量机特征选择和参数联合优化算法 [J]. 计算机应用，2016, 36 (5): 1330-1335. (Zhang Jin, Ding Sheng, Li Bo. Improved particle swarm optimization algorithm for support vector machine feature selection and optimization of parameters [J]. Journal of Computer Applications, 2016, 36 (5): 1330-1335.)

[20] Dash M, Choi K, Scheuermann P, *et al.* Feature Selection for Clustering-A Filter Solution [C]// Proc of IEEE International Conference on Data Mining. 2002: 115-122.